(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 January 2005 (13.01.2005)

PCT

(10) International Publication Number
WO 2005/003308 A2

(51) International Patent Classification⁷: C12N

(21) International Application Number: PCT/US2004/019932

(22) International Filing Date: 22 June 2004 (22.06.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/482,420 25 June 2003 (25.06.2003) US

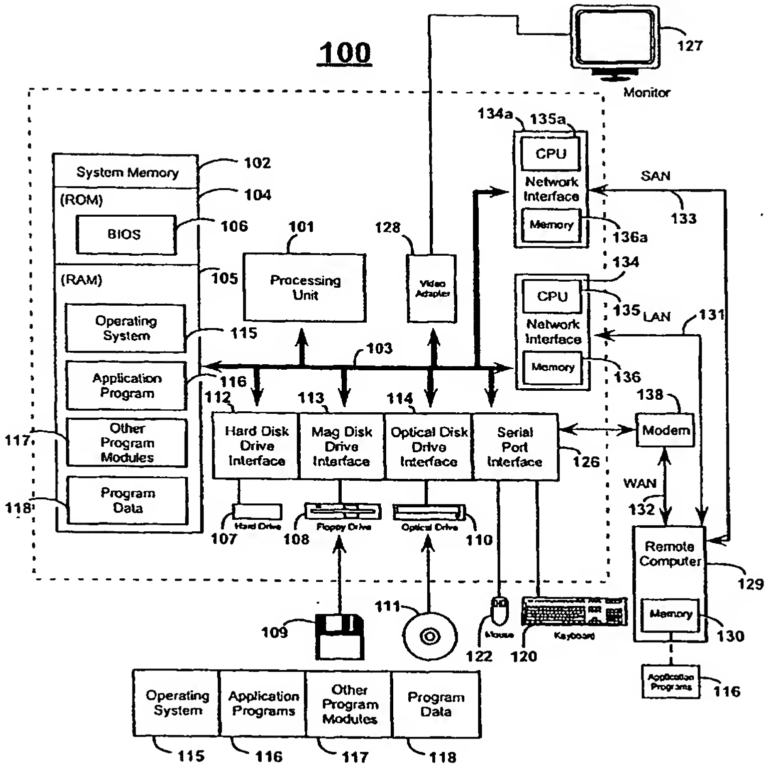
(71) Applicant *(for all designated States except US)*: **SMITHKLINE BEECHAM CORPORATION** [US/US]; One Franklin Plaza, Philadelphia, PA 19101 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **AGARWAL, Pankaj** [IN/US]; GlaxoSmithKline, 709 Swedeland Road, King of Prussia, PA 19406 (US). **HURLE, Mark, Robert** [US/US]; GlaxoSmithKline, 709 Swedeland Road, King of Prussia, PA 19406 (US). **KABNICK, Karen, Stephanie** [US/US]; GlaxoSmithKline, 709 Swedeland Road, King of Prussia, PA 19406 (US). **LIU, Liwen** [CN/US]; GlaxoSmithKline, Five Moore Drive, PO Box 13398, Research Triangle Park, NC 27709 (US). **MAGID-SLAV, Michal** [IL/US]; GlaxoSmithKline, 709 Swedeland Road, King of Prussia, PA 19406 (US). **MCALLISTER, Paul, Robert** [US/US]; GlaxoSmithKline, 709 Swedeland Road, King of Prussia, PA 19406 (US). **SEARLS, David, Burdette** [US/US]; GlaxoSmithKline, 709 Swedeland, King of Prussia, PA 19406 (US). **TATSUOKA, Kay, Satoshi** [US/US]; GlaxoSmithKline, 709 Swedeland Road, King of Prussia, PA 19406 (US). **ZAYKIN, Dmitri, V** [RU/US]; GlaxoSmithKline, Five Moore Drive, PO Box 13398, Research Triangle Park, NC 27709 (US).

(74) Agents: **LEVY, David, J.** et al.; GlaxoSmithKline, Five Moore Drive, PO Box 13398, Research Triangle Park, NC 27709 (US).

[Continued on next page]

(54) Title: BIOLOGICAL DATA SET COMPARISON METHOD



(57) Abstract: A method of identifying a relationship between a set of one or more candidate biomolecules and a set of one or more reference biomolecules, the method including inputting to a computer a query set describing the one or more candidate biomolecules; comparing the query set with a target database describing the one or more reference biomolecules wherein the one or more reference biomolecules grouped into one or more buckets and wherein the one or more reference biomolecules of each bucket share a common property; counting a number of matches between each query set and each buckets of the target database; and statistically analyzing the number of matches to each bucket wherein the presence of a statistically significant match identifies a relationship between a the query set and a bucket of the target database.

WO 2005/003308 A2